

# Detecting community structure in complex networks via node similarity

Ying Pan<sup>a,b,\*</sup>, De-Hua Li<sup>a</sup>, Jian-Guo Liu<sup>c,d</sup>, Jing-Zhang Liang<sup>b</sup>

<sup>a</sup> Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China

<sup>b</sup> Information Network Center, Guangxi University, Nanning 530004, China

<sup>c</sup> Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China

<sup>d</sup> Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700, Fribourg, Switzerland

The detection of the community structure in networks is beneficial to understand the network structure and to analyze the network properties. Based on node similarity, a fast and efficient method for detecting community structure is proposed, which discovers the community structure by iteratively incorporating the community containing a node with the communities that contain the nodes with maximum similarity to this node to form a new community. The presented method has low computational complexity because of requiring only the local information of the network, and it does not need any prior knowledge about the communities and its detection results are robust on the selection of the initial node. Some real-world and computer-generated networks are used to evaluate the performance of the presented method. The simulation results demonstrate that this method is efficient to detect community structure in complex networks, and the ZLZ metrics used in the proposed method is the most suitable one among local indices in community detection.

## 1. Introduction

Recently, complex networks have attracted considerable attention in many fields for representation of a variety of complex systems, such as biological and social systems, the Internet, the World Wide Web, and so on [1–4]. Community structure is an important property of complex networks, which is the tendency for nodes to divide into groups, with dense connections within groups and only sparse connections between them [5]. It is significant to analyze community structure because it often corresponds to functional units such as cycles or pathways in metabolic networks or collections of pages on a single topic on the web. Moreover, a number of recent results suggest that networks have properties at the community level that are quite different from their properties at the level of the entire networks and ignoring community structure may miss many interesting features [6].

A large number of methods have been developed to detect community structure in networks in the past years. There are mainly two kinds of clustering algorithms, one is partitioning algorithm, and the other is the hierarchical clustering method. The Kernighan–Lin algorithm [7], which uses a greedy algorithm to optimize the value of the edges within community minus those between community, and the spectral bisection algorithm [8], which is based on the eigenvectors of the Laplacian matrix of graph, are two classical algorithms related to the ideas of graph partitioning in graph theory and computer science. They can find the community structure efficiently in the networks in the case that the number of communities in the

\* Corresponding author at: Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China. Tel.: +86 27 87544769; fax: +86 27 87547408.

E-mail addresses: panyingpan@gmail.com (Y. Pan), liujg004@ustc.edu.cn (J.-G. Liu).

networks is given before. The agglomerative and divisive methods are two types of hierarchical clustering algorithms in sociology to find community structure in networks. They first compute the intensity of link between each pair nodes based on different methods, such as edge betweenness [5,9], edge clustering coefficient [10], dissimilarity index [11], information centrality [12], similarity based on random walks [13], clustering centrality [14], and so on. Then, by repeatedly incorporating the two nodes with the highest intensity of link (agglomerative method), or repeatedly removing the edge with the lowest intensity (divisive methods), the partition results of the networks are obtained. In 2002, Girvan and Newman proposed a divisive method that based on the concept of edge betweenness to identify the community structure of the network [5]. Although this method has been successfully applied to a variety of networks, the complexity of it is not ideal, running in  $O(m^2n)$  time on an arbitrary network with  $m$  edges and  $n$  nodes, or  $O(n^3)$  time on a sparse network (a network with  $m \sim n$ , which covers most real-world networks of interest). To decrease the complexity, Newman [15] proposed a fast algorithm in 2004 for detecting community structure based on the idea of modularity [9], which runs in time  $O((m+n)n)$  on an arbitrary network and  $O(n^2)$  for sparse network. Moreover, Clauset et al. [16] presented a hierarchical agglomeration algorithm to detect the community in very large networks, which performed the same greedy optimization of Ref. [15] yet adopted more sophisticated data structures to give out identical results, in running time  $O(md \log n)$ , where  $d$  is the depth of the dendrogram describing the community structure. For a sparse and hierarchical network with  $m \sim n$  and  $d \sim \log n$ , this method has near linear running time  $O(n \log^2 n)$ . Another near linear time algorithm is proposed in Ref. [17], which based on label propagation and used only local information to analyze community structures, in time  $O(m+n)$  for large-scale complex networks. In Ref. [18], Xiang et al. designed a faster algorithm that based on subgraph similarity to identify community, which provided the same level of reliability while takes shorter time than the algorithm proposed by Clauset et al. [16].

In this paper, we propose an agglomerative community detection method based on node similarity. Each node is initially considered as a community and then starting with an arbitrary node, the community of this node incorporates with those communities which contain the maximum similarity nodes to form a new community, performing the incorporation repeatedly until all nodes in the network have been visited. As we will show, the method has some characteristics different from other techniques, including it does not need any prior knowledge about the number of communities in the networks, and its detection results are not influenced by the initial node. Since the proposed method requires only the local information of the nodes, its running time is much lower than many other methods, with computational complexity of  $O(nk)$  on an arbitrary network, where  $k$  is the mean node degree of the whole network.

The rest of this paper is organized as follows. First the concept and computation method of node similarity are briefly introduced in Section 2. Then the proposed method is described detailedly in Section 3 and the simulation results are presented in Section 4. Finally, the conclusions and comments are drawn in Section 5.

## 2. Node similarity

The similarity is a measure of closeness between a pair of nodes. Several node similarity metrics on basis of local information are described in Ref. [19], which show different performance for detecting community structure in complex networks. And the discussion in Appendix reveals that the similarity metric proposed by Zhou et al. (Zhou-Lü-Zhang Index) [19] is the most suitable one to detect community structure in complex networks. This similarity metric, first proposed in the Ref. [20], is motivated by the resource allocation process taking place in networks and has been successfully applied in the link prediction of transportation networks [19] and personal recommendation [21–23], which are all based on the ZLZ Index to measure the object and user similarity in the bipartite networks.

Consider a pair of nodes  $i$  and  $j$ , and the node  $i$  can send some resource to  $j$ , with their common neighbors playing the role of transmitters. Assuming that each transmitter has a unit of resource, then the amount of resource  $j$  received, namely the similarity between nodes  $i$  and  $j$  is [19]:

$$S_{ij} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{k(z)}, \quad (1)$$

where  $\Gamma(i)$  is the set of neighbors of  $i$ , and  $z \in \Gamma(i) \cap \Gamma(j)$  are the common neighbors of node  $i$  and  $j$ , and  $k(z)$  is the degree of node  $z$ .

So we can measure the closeness of each pair of nodes according to Eq. (1). However, the metric can not differentiate the tightness relation between a pair of nodes whether they are connected directly or indirectly, which may result in inaccurate detection for community structure on the networks. For example, as for the Zachary's karate club network [24] in Fig. 3, according to Eq. (1), the similarity between nodes 3 and 34 is the maximum among the similarities between node 3 and the others, and then they would be group into a community in the proposed algorithm (the community detection procedure can be found in Section 3.1). Obviously, the pair of nodes does not belong to the same community in fact. Therefore, for a pair of nodes, in the cases of connected directly or indirectly, the similarities between them are different for detecting community structure on the networks. Thus we calculate the similarity between  $i$  and  $j$  as follows:

$$S_{ij} = \begin{cases} \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{k(z)} & \text{if } i, j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

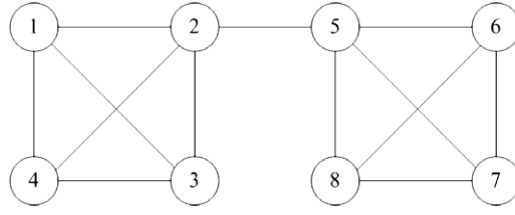


Fig. 1. The example network.

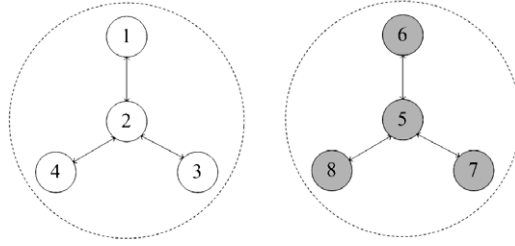


Fig. 2. Detecting community structure in the example network using the proposed method. Two communities are detected, labeled by shaded and open circles respectively.

### 3. The proposed method

In this section, the principle and implementation of the proposed method is detailed, and then its computational complexity is analyzed.

#### 3.1. Community detection

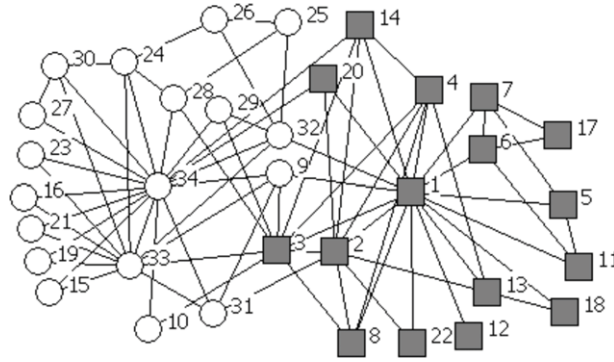
The proposed method is based on the idea that two nodes should have more chance to belong to the same community when their similarity is larger.

For a network with  $n$  nodes, we first calculate the similarity of each pair of nodes according to Eq. (2) and thus obtain a  $n \times n$  similarity matrix  $S = \{s_{ij}\}$ , where the element  $s_{ij}$  denote the similarity between nodes  $i$  and  $j$ . For convenience, each row of the similarity matrix is stored as a linked list in our algorithm, where the head node of linked list save the row number of the similarity matrix, and the other nodes of linked list save the order number of its neighbors with maximum similarity (including the case of multi-neighbors with the same maximum similarity). Then, the procedure of our method is as follows:

- (i) Assume that each node is considered as a community, and select an arbitrary node as the initial node.
- (ii) Incorporating the community containing this node with the communities that contain the nodes with maximum similarity to this node to form a new community.
- (iii) Determine the next node to be processed. Choose the node with maximum similarity to the current node as the next node. If this new node is not contained in the current community, go to step (ii) to perform further incorporation. Otherwise, randomly choose a new node that is not visited as the initial node and go to step (ii) to perform a new community incorporation.
- (iiii) Repeat the step (ii) and step (iii) until all nodes in the networks have been visited.

To make our method clear to readers, we show a small scale example network consisted of eight nodes, as shown in Fig. 1. According to Eq. (2), the similarities of the nodes in Fig. 1 are computed and those with maximum similarity are listed in Table 1. Using the results presented in Table 1, Fig. 2 illustrates the procedure of community incorporations, where the arrows point to the neighbors with maximum similarity to the node that locates at the other end of the arrows. In addition, the arrows also represent that the communities containing the nodes positioned at the two ends of the arrows should be incorporated into a community. After the incorporations denoted by all the arrows are finished, the detection results are obtained. Two communities are detected by the proposed algorithm for the example network, denoted by shaded and open circles in Fig. 2, respectively.

In the above steps, a special situation should be considered. If a node and its neighbors have not any common neighbors, that is, the similarities value of a node and its neighbors are all equal to zero, then the node selects the communities that contain its neighbors with maximum degree to join. The reason behind the selection is from the preferential attachment mechanism proposed by Barabási and Albert in scale-free evolving network [25], which the new node to a network tends to connect the node with large degree. By analogy, we believe that the node tends to group with its neighbor node with maximum degree in the case of no common neighbors of them. Furthermore, to improve the quality of the proposed method



**Fig. 3.** The network of friendships between individuals in the karate club study of Zachary. The administrator and the instructor are represented by nodes 1 and 33 respectively. Shaded squares represent individuals to who ended up aligning with the club's administrator after the fission of the club, open circles those who aligned with the instructor.

we introduce the similarity threshold  $\varepsilon$ , that is, if the ratio of similarities between the  $i$ th node with other two nodes is lower than the similarity threshold, we can consider the similarities between the  $i$ th node with these two nodes is approximate equivalent (surely, the similarity threshold value is different for different network, and one can tune the value to obtain the highest detection accuracy).

### 3.2. Computing complexity analysis

Obviously, the computational cost of the proposed algorithm mainly consists of three operations: computations of the similarities, searching the next node to be processed, and incorporations of communities. Since the next node to be processed is the node with maximum similarity to the current node, which is stored in the linked list of this node in the position next to the head node and can be obtained directly, the time of searching it can be neglected. As a consequence, the cost of the developed method is basically determined by the operations of computations of the similarities and incorporations of communities.

The similarity measure requires only the information of the nearest neighbors, so it requires very low computational cost. Since the computational cost of the Eq. (2) is  $O(1)$ , then the computations of the similarities of a node with its  $k$  neighbors are  $O(k)$ , where  $k$  is the mean node degree of the whole network. So the similarity measure of the network with  $n$  nodes takes an approximate running time of  $O(nk)$ . It is easy to observe that the incorporation of the community containing the current node and the communities that contain the nodes with maximum similarity takes  $O(1)$  CPU time, therefore the computational complexity of the incorporations of communities for a network with  $n$  nodes is  $O(n)$ . According to above analysis, the total computational complexity of the proposed method is  $O(nk + n) \sim O(nk)$ .

Beside the time complexity, the proposed method also required relatively less memory space. In the worst case that the similarities of a node and its neighbors are the same value, the memory of the link list of the node required is of the order  $O(k)$ . So the whole memory of the network with  $n$  nodes require  $O(nk)$ . It is clear that the proposed method is time efficient and required relatively less memory.

## 4. Simulation results

To evaluate the performance of the proposed method, some networks, such as the Zachary's karate club network, football network, dolphin association network and the computer-generated networks, are used to be the test networks. And all experiments are run on a PC with 2.0 GHz processor and 2.0 GB memory.

### 4.1. Zachary's karate club network

The Zachary's karate club network [24] is one of the classic studies in social network analysis and has been used as one of the typical test examples by many researchers to detect community structures in complex network [9,15,17]. The club network consists of 34 member nodes, and splits in two smaller clubs after a dispute arose during the course of Zachary's study. In Fig. 3 we show a consensus network structure extracted from Zachary's observations.

The similarities of the nodes in Fig. 3 are computed according to Eq. (2) and those with maximum similarity are listed in Table 2 (the similarity threshold  $\varepsilon$  is set 0.9). And the Fig. 4 presents the procedure of community incorporations. Two communities are detected by the proposed algorithm within less than 5 ms CPU time, denoted by shaded and open circles in Fig. 4, respectively. It is seen that the partition using the proposed algorithm is consistent with the actual division of original club. The algorithm proposed by Clauset, Newman and Moore (CNM) [16] also detects community on this network with less than 5 ms, but it is noted that it classifies four instead of two communities, which splits the larger one into three sub-graphs (the program code for the CNM algorithm is directly download from the personal homepage of Clauset).

**Table 1**

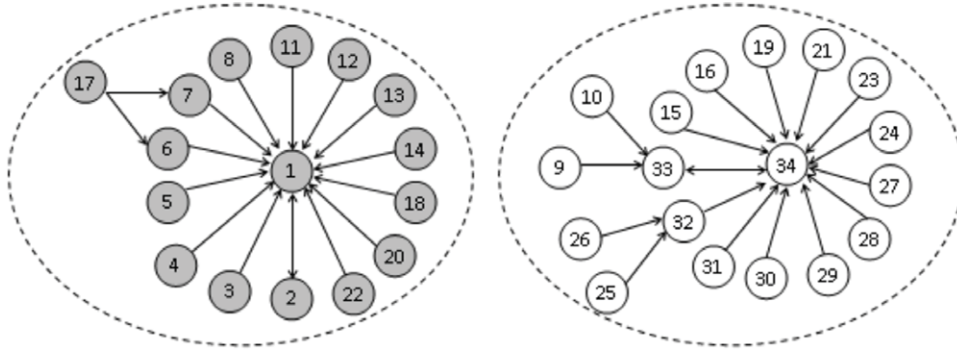
Nodes and their neighbors with maximum similarity (MSNs) of the example network.

Node	MSNs	Node	MSNs	Node	MSNs
1	2	4	2	7	5
2	1, 3, 4	5	6, 7, 8	8	5
3	2	6	5		

**Table 2**

Nodes and their neighbors with maximum similarity (MSNs) of Zachary's network.

Node	MSNs	Node	MSNs	Node	MSNs
1	2	13	1	24	34
2	1	14	1	25	32
3	1	15	34	26	32
4	1	16	34	27	34
5	1	17	6, 7	28	34
6	1	18	1	29	34
7	1	19	34	30	34
8	1	20	1	31	34
9	33	21	34	32	34
10	33	22	1	33	34
11	1	23	34	34	33
12	1				

**Fig. 4.** Detecting community structure in the Zachary's network using the proposed method. Two communities are detected, labeled by shaded and open circles respectively.

#### 4.2. College football network

College football network [5] represents the schedule of games between American college football teams in a single season. The network constructed by 115 teams is divided into 12 groups or “conferences”, with intraconference games being more frequent than interconference games.

The procedure of incorporations of the community containing current node with the communities that contain the nodes with maximum similarity is shown in Fig. 5 ( $\varepsilon = 0.75$ ). As Fig. 5 reveals, 109 of 115 teams are classified correctly and eleven communities are detected by the proposed method. Seven communities are detected correctly, including Atlantic Coast, Big Ten, Big Twelve, Conference USA, Mountain West, Pacific Ten and Southeastern. Although the Boise State team is a member of Western Athletic, it has higher similarity with the members of Sun Belt than other members, so it is assigns into Sun Belt in our algorithm.

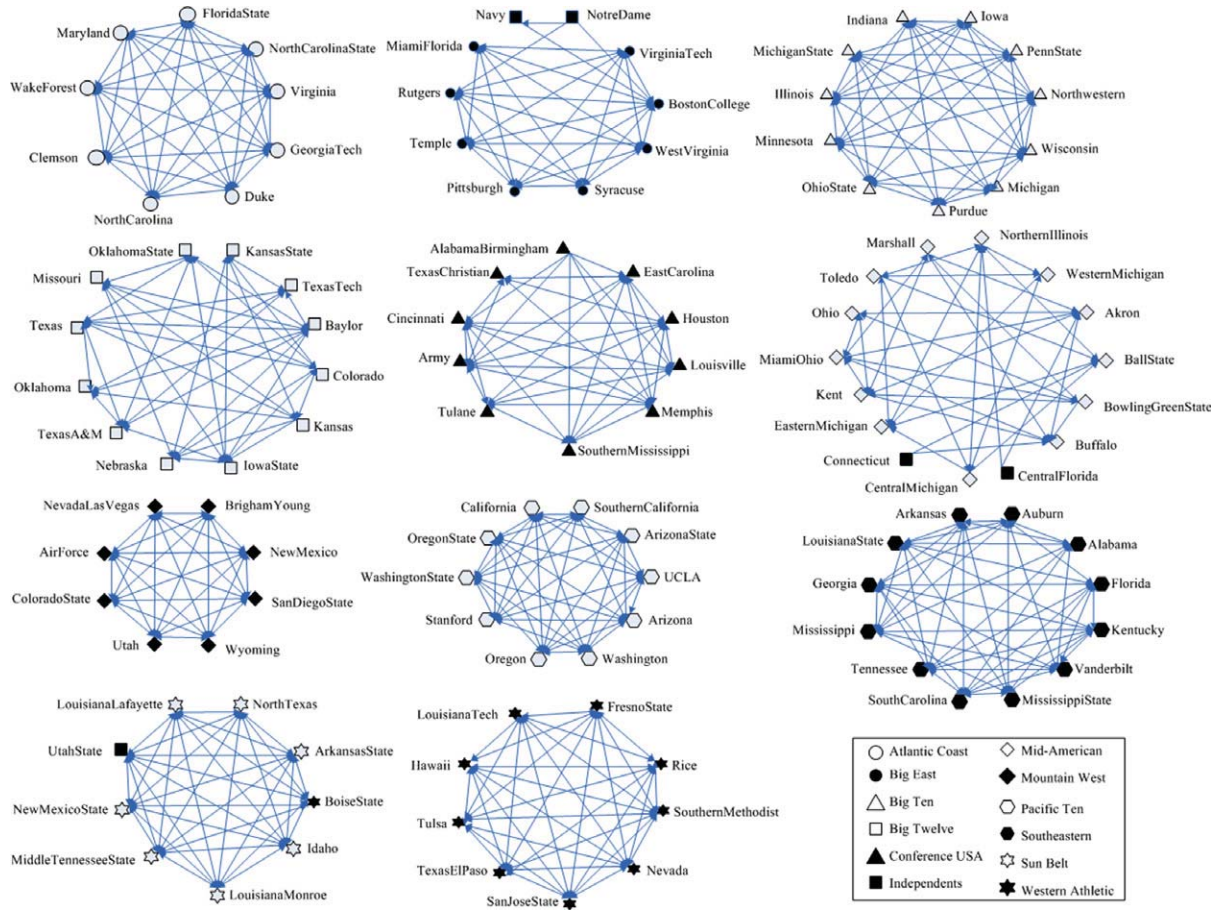
The computing time of the proposed method to completion in the network is about 30 ms, a litter faster than the CNM algorithm [16], which detected seven communities with 48 ms. Moreover, only four fully correct communities are detected by the CNM algorithm.

#### 4.3. Associations between 62 dolphins

The network describes the associations between 62 dolphins living in Doubtful Sound, New Zealand, compiled by Lusseau [26] from seven years of field studies of the dolphins. As Lusseau points out [27], the network splits naturally into two large groups, which correspond to a known division of the dolphin community, and the larger of the two also splits into smaller subgroups.

The largest subgroup consists almost of females and the others almost males. For this network, four communities are detected by the proposed algorithm within less than 5 ms CPU run time. The largest community, represented by blue





**Fig. 5.** Detecting community structure in the College football network using the proposed method. Eleven communities are detected. The real-world communities are denoted by the different shapes as indicated in the legend.

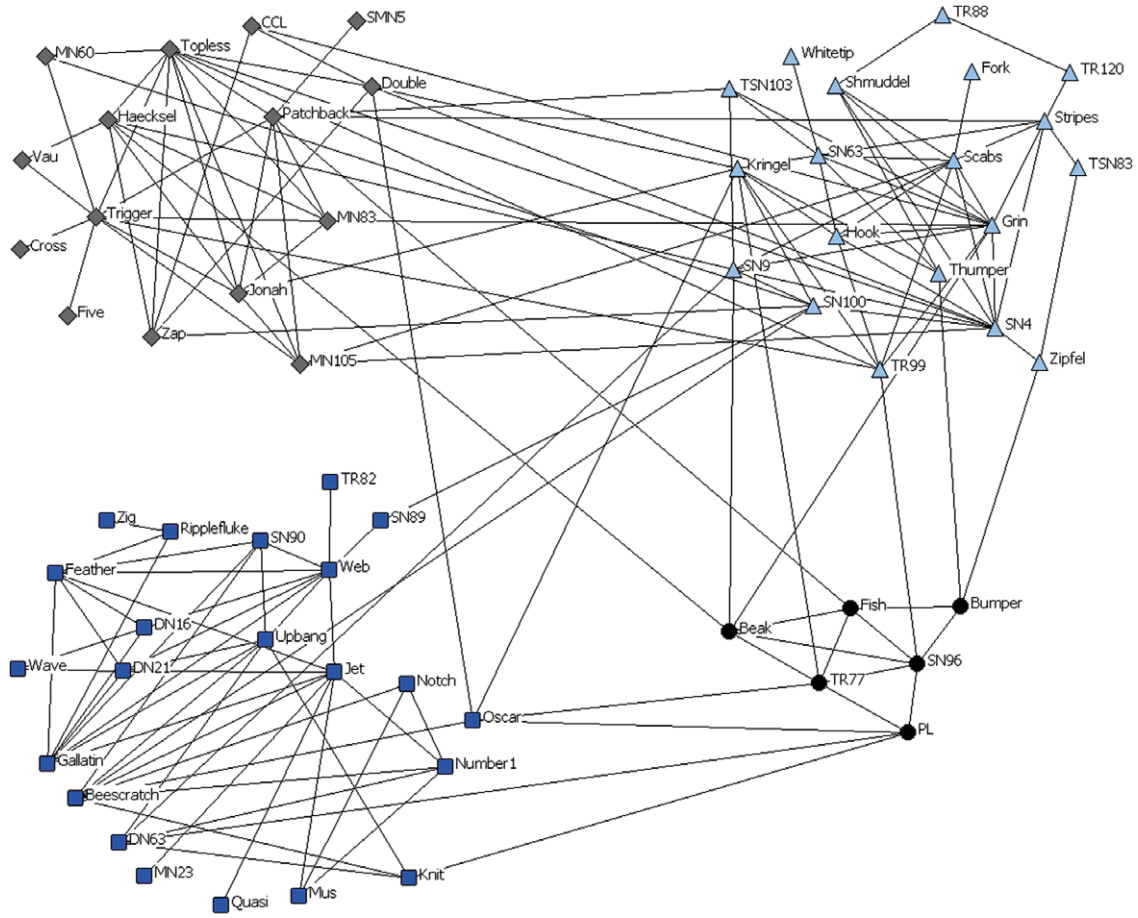
squares, has fewest relations with other communities. The two communities, denoted by shaded diamond and blue triangle respectively, have most intercommunity relations, as shown in Fig. 6 ( $\varepsilon = 0.75$ ).

#### 4.4. Computer-generated networks

The presented method is applied to the class of benchmark computer-generated networks proposed by Lancichinetti, Fortunato and Rachicchi [28]. Such networks have a heterogeneous distribution of node degree and community size, which are features of real networks, so it can be considered a proxy of a real network with community structure and appropriate to evaluate the performance of the community detection algorithms. In the benchmark networks, both the degree and the community size distributions are power laws, with exponents  $\gamma$  and  $\beta$ , respectively, and the number of nodes is  $N$  and the average degree is  $\langle k \rangle$ . Each node shares a fraction  $1 - \mu$  of its links with the other nodes of its community and a fraction  $\mu$  with the other nodes of the network, calling  $\mu$  the mixing parameter of the network.

For evaluating the performance of the community identification, a metric should be adopted. Although the modularity proposed in Ref. [9] is the most popular metric for community identification and the results according to the maximal modularity looks reasonable, it has an intrinsic limit that makes small communities hard to detect [29]. So an alternative measure, namely normalized mutual information [30], is used to quality the performance of the proposed algorithm in the paper, which is a measure of similarity of partitions borrowed from information theory and has proved to be reliable.

In Fig. 7 we show the variation of the normal mutual information obtained by the proposed method on the benchmark networks, with the mixing parameter  $\mu$  from 0.1 to 0.6,  $\langle k \rangle = 20$ ,  $\gamma = 2.5$  and  $\beta = 1.5$  (the similarity threshold is set  $\varepsilon = 0.75$ ). The four panels in Fig. 7 correspond to four types of networks with number of nodes is  $N = 1000, 5000, 10\,000$  and  $100\,000$ , respectively. We have chosen larger and larger size of network in order to check how the performance is affected by the network size. With  $\mu$  increasing, the community structure in the networks becomes more fuzzy and harder to identify, and  $\mu = 0.5$  (dashed vertical line in Fig. 7) marks the border beyond which community are no longer defined in the strong sense, i.e., such that each node has more neighbors in its own community than in the others. Inspection of Fig. 7 reveals that the proposed method performs very well, though the curves goes down more quickly for larger networks. The average



**Fig. 6.** Detecting community structure in the dolphins association network using the proposed method. Four communities are detected, which denoted by the different shapes.

values of the normalized mutual information of our method are 0.997, 0.978, 0.966 and 0.939 when  $\mu \leq 0.5$  for  $N = 1000, 5000, 10\,000$  and  $100\,000$ , respectively. On the same plot we also show the performance of the CNM algorithm [16]. As we can see, our algorithm performs better than that algorithm, which average values of the normalized mutual information are 0.985, 0.962, 0.950 and 0.916 when  $\mu \leq 0.5$  for  $N = 1000, 5000, 10\,000$  and  $100\,000$ , respectively. Moreover, the CPU time of our method is less than the CNM algorithm on the benchmark networks, i.e., our method uses about 0.5 s, 2 s, 10 s and 90 s, but the CNM algorithm uses nearly 1 s, 8 s, 20 s and 150 s on the network with  $N = 1000, 5000, 10\,000$  and  $100\,000$ , respectively.

## 5. Conclusions

Based on the node similarity, a fast and efficient algorithm for detecting community structure in networks is developed, which ensures that a pair of nodes with higher similarity has more likely to be grouped into a community. It doesn't need to know the structure of the whole network in advance and has lower computational complexity than other methods. The algorithm has been applied to a variety of networks, both real and artificial, showing that it is rather efficient to discover the community structure of complex networks.

## Appendix. Comparing similarity measures

The aim of this section is to investigate the problem of which node similarity metric is more suitable to detect community structure in complex networks. Here we concentrate on ten similarity metrics on basis of local information because they require only local information, thereby needing less computational time, including Common Neighbors, Salton index [31], Jaccard Index [32], Sørensen Index [33], Hub Prompted Index [34], Hub Depressed Index [19], Leicht–Holme–Newman Index [35], Preferential Attachment [36], Adamic–Adar Index [37] and Zhou–Lü–Zhang Index [19]. In Tables 3 and 4, we respectively present the normalized mutual information and the CPU times of the similarity metrics on the real and

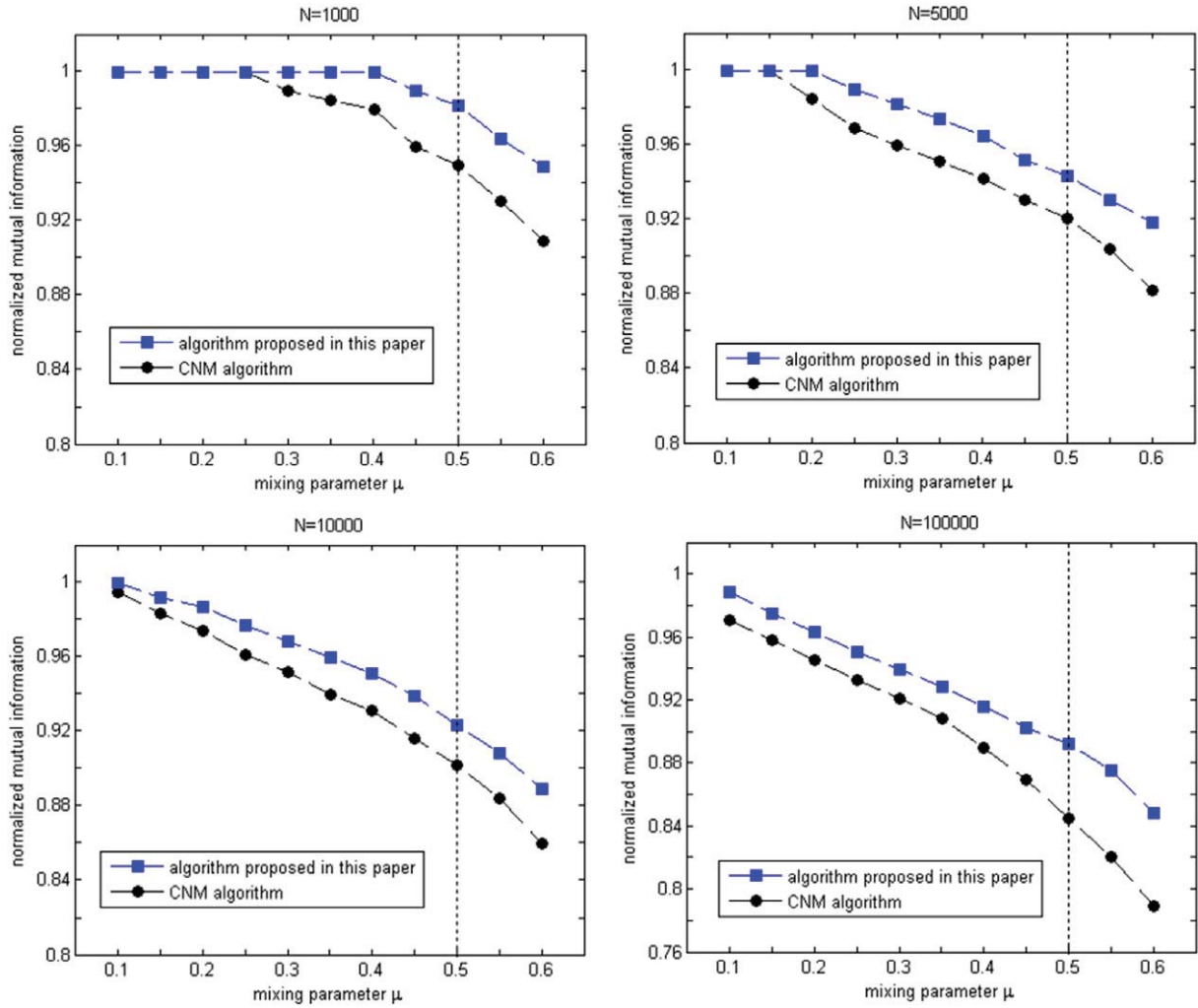


Fig. 7. Comparison of the normalized mutual information between the proposed algorithm (circle) and the CNM algorithm [16] (square) on the benchmark computer-generated network [24].

Table 3

Normalized mutual information of ten similarity metrics (the abbreviations, CN, Salton, Jaccard, Sørensen, HPI, HDI, LHN, PA, AA and ZLZ stand for Common Neighbors, Salton Index, Jaccard Index, Sørensen Index, Hub Prompted Index, Hub Depressed Index, Leicht–Holme–Newman Index, Preferential Attachment, Adamic–Adar Index and Zhou–Lü–Zhang Index, respectively.)

Similarity metrics	Real network		Benchmark network with $\langle k \rangle = 20$ , $\gamma = 2.5$ , $\beta = 1.5$ and $\mu = 0.2$			
	Karate club network	Football network	$N = 1000$	$N = 5000$	$N = 10\,000$	$N = 100\,000$
CN	1	0.865	1	0.997	0.971	0.953
Salton	1	0.922	0.992	0.983	0.961	0.928
Jaccard	0.906	0.885	0.970	0.914	0.887	0.864
Sørensen	0.914	0.938	0.979	0.937	0.901	0.877
HPI	1	0.887	1	1	0.977	0.955
HDI	0.890	0.938	0.870	0.861	0.830	0.802
LHN	0.945	0.922	0.995	0.993	0.973	0.949
PA	0.947	0.922	0.812	0.805	0.776	0.744
AA	1	0.946	1	1	0.988	0.968
ZLZ	1	0.946	1	1	0.985	0.960

benchmark networks (described on Section 4). From Table 3, we can find that both Zhou–Lü–Zhang Index and Adamic–Adar Index have good performance evaluated by the normalized mutual information although the former is a little worse than the latter only when the size of network gets large. However, the CPU times of Zhou–Lü–Zhang Index uses are much less than Adamic–Adar Index, as shown in Table 4. Therefore, the Zhou–Lü–Zhang Index is the most suitable one among the ten similarity metrics to detect community structure in complex networks.



**Table 4**

CPU times of ten similarity metrics (in millisecond).

Similarity metrics	Real network		Benchmark network with $\langle k \rangle = 20$ , $\gamma = 2.5$ , $\beta = 1.5$ and $\mu = 0.2$			
	Karate club network	Football network	$N = 1000$	$N = 5000$	$N = 10\,000$	$N = 100\,000$
CN	16	30	514	1975	10855	101449
Salton	4	77	542	2395	11542	125470
Jaccard	20	30	955	8786	45147	510854
Sørensen	5	28	542	2371	14172	145532
HPI	4	31	548	2427	13421	102483
HDI	4	16	523	2313	13874	99856
LHN	20	27	531	2336	11210	154892
PA	4	33	511	1942	10854	100142
AA	21	187	4258	16734	97856	843677
ZLZ	5	30	505	1950	10114	98380

## References

- [1] S.H. Strogatz, Exploring complex networks, *Nature* 410 (2001) 268–276.
- [2] R. Albert, A.L. Barabasi, Statistical mechanics of complex networks, *Rev. Modern. Phys.* 74 (2002) 47–97.
- [3] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks, *Adv. Phys.* 51 (2002) 1079–1187.
- [4] M.E.J. Newman, The structure and function of complex networks, *SIAM Rev.* 45 (2003) 167–256.
- [5] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, *Proc. Natl. Acad. Sci. USA* 99 (2002) 7821–7826.
- [6] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, *Phys. Rev. E* 74 (2006) 036104.
- [7] B.W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, *Bell. Syst. Tech. J.* 49 (1970) 291–307.
- [8] M. Fiedler, Algebraic connectivity of graphs, *Czech Math. J.* 23 (1973) 298–305.
- [9] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E* 69 (2004) 026113.
- [10] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, *Proc. Natl. Acad. Sci. USA* 101 (2004) 2658–2663.
- [11] H. Zhou, Distance, dissimilarity index, and network community structure, *Phys. Rev. E* 67 (2003) 061901.
- [12] S. Fortunato, V. Latora, M. Marchiori, A method to find community structures based on information centrality, *Phys. Rev. E* 70 (2004) 056104.
- [13] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: *Proc. of the 20th International Symposium on Computer and Information Sciences*, Springer-Verlag, Berlin, 2005, pp. 284–293.
- [14] B. Yang, J. Liu, Discovering global network communities based on local centralities, *ACM Trans. on the Web* 2 (1) (2008) 1–32.
- [15] M.E.J. Newman, Fast algorithm for detecting community structure in networks, *Phys. Rev. E* 69 (2004) 066133.
- [16] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, *Phys. Rev. E* 70 (2004) 066111.
- [17] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, *Phys. Rev. E* 76 (2007) 036106.
- [18] B. Xiang, E.H. Chen, T. Zhou, Finding community structure based on subgraph similarity, *Stud. Comput. Intell.* 207 (2009) 73–81.
- [19] T. Zhou, L. Lü, Y.C. Zhang, Predicting missing links via local information, *Eur. Phys. J. B* 71 (2009) 623–630.
- [20] T. Zhou, J. Ren, M. Medo, Y.C. Zhang, Bipartite network projection and personal recommendation, *Phys. Rev. E* 76 (2007) 046115.
- [21] J.-G. Liu, B.H. Wang, Q. Guo, Improved collaborative filtering algorithm via information transformation, *Int. J. Modern Phys. C* 20 (2009) 285–293.
- [22] J.-G. Liu, T. Zhou, B.H. Wang, Y.C. Zhang, Q. Guo, Effect of user tastes on personalized recommendation, *Int. J. Modern Phys. C* 20 (12) (2009) 1925–1932.
- [23] J.-G. Liu, T. Zhou, H.A. Che, B.H. Wang, Y.C. Zhang, Effects of high-order correlations on personalized recommendations for bipartite networks, *Physica A* 389 (2010) 881–886.
- [24] W.W. Zachary, An information flow model for conflict and fission in small groups, *J. Anthropol. Res.* 33 (1977) 452–473.
- [25] A.L. Barabási, R. Albert, Emergence of scaling in random networks, *Science* 286 (1999) 509–512.
- [26] D. Lusseau, The emergent properties of a dolphin social network, *Proc. R. Soc. Lond. B* 270 (suppl.) (2003) S186–S188.
- [27] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, *Behav. Ecol. Sociobiol.* 54 (2003) 396–405.
- [28] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, *Phys. Rev. E* 78 (2008) 046110.
- [29] S. Fortunato, M. Barthélemy, Resolution limit in community detection, *Proc. Natl. Acad. Sci. U.S.A.* 104 (2007) 36.
- [30] L. Danon, A. Díaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, *J. Stat. Mech.: Theory Exp.* (2005) P09008.
- [31] G. Salton, M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA, 1986.
- [32] P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bull. Soc. Vaudoise Sci. Natur.* 37 (1901) 547–579.
- [33] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, *Vidensk Selsk. Biol. Skr.* 5 (1948) 1–34.
- [34] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.L. Barabasi, Hierarchical organization of modularity in metabolic networks, *Science* 297 (2002) 1551–1555.
- [35] E.A. Leicht, P. Holme, M.E.J. Newman, Vertex similarity in networks, *Phys. Rev. E* 73 (2006) 026120.
- [36] Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative filtering, in: *Proceedings of the 5th ACM/IEEECS Joint Conference on Digital Libraries*, ACM Press, New York, NY, USA, 2005, pp. 141–142.
- [37] L.A. Adamic, E. Adar, Friends and neighbors on the web, *Soc. Netw.* 25 (2003) 211–230.